electronics

MDPI

Article

Design and Evaluation of Knowledge-Distilled LLM for Improving the Efficiency of School Administrative Document Processing

Younhee Hong

Semiconductor Specialized Graduate School Project Group, Chungnam National University, Daejeon 34134, Republic of Korea; yhhong@mokwon.ac.kr

Abstract

This study proposed OP-LLM-SA, a knowledge distillation-based lightweight model, for building an on-premise AI system for public documents, and evaluated its performance based on 80 public documents. The token accuracy was 92.36%, and the complete sentence rate was 97.19%, showing meaningful results compared to the original documents. During inference, the GPU environment required only about 4.5 GB, indicating that the model can be used on general office computers, and Llama-3.2's Korean language support model showed the best performance among the LLMs. This study is significant in that it proposes a system that can efficiently process public documents in an on-premise environment. In particular, it is expected to be helpful for teachers who are burdened with processing public documents. In the future, we plan to conduct research to expand the scope of application of text mining technology to various administrative document processing environments that handle public documents and personal information, as well as school administration.

Keywords: knowledge distillation; model compression; administrative document processing; on-premise AI; natural language processing (NLP)

Academic Editor: Alexander Gegov

Received: 19 August 2025
Revised: 23 September 2025
Accepted: 26 September 2025
Published: 29 September 2025

Citation: Hong, Y. Design and Evaluation of Knowledge-Distilled LLM for Improving the Efficiency of School Administrative Document Processing. *Electronics* **2025**, *14*, 3860. https://doi.org/10.3390/electronics14193860

1. Introduction

Most administrative processing and decision making in public institutions is carried out through a formal document system. Public documents are official documents used by public institutions to express the basis and actions of decision making and work in writing. These documents record the results of public actions and serve as a means of administrative communication for the exchange of information between institutions and citizens [1]. Public documents, which are such an important medium, require considerable time and effort to prepare. In the administration of South Korea's schools, teachers are responsible for creating and managing official documents, placing a heavy administrative burden on them [2]. Teachers handle many administrative tasks in addition to their educational activities, and excessive paperwork is a major contributor to their workload. Therefore, there is a pressing need to establish an efficient document creation and processing system [3].

Large language models are considered an effective alternative for efficient document management and intelligent document processing [4]. In particular, natural language processing technology, through LLMs, extracts content and information and constructs new words and contexts based on the extracted information [5]. Such natural language processing technology is considered to improve the usability of systems and enable efficient technical support through data consistency [6].

The documents generated in schools' administration are becoming more voluminous, and the need for effective management is being emphasized, but there are limitations to directly applying large language models. School administration data mostly consists of students' personal information, and there is a possibility of sensitive information being learned or leaked during LLM data learning [7]. In particular, schools in South Korea are classified as public institutions and are subject to the information protection management system for public institutions in accordance with Article 32-2 of the Personal Information Protection Act. Accordingly, school administration data is recognized as public data, and schools have established management systems to ensure that personal information is strongly protected, such as access control, which limits the use of general cloud LLM services [8].

In order to introduce LLMs into school administration systems, it is essential to build an On-Premise LLM environment utilizing the school's internal servers. An LLM is composed of countless parameters and requires high-performance GPUs, making it unsuitable for a school's internal servers. Therefore, it is necessary to reduce the size of the model so that it can be executed with limited memory. Reducing the size of the model allows it to be executed even on low-end servers, reducing the cost of the construction of a school internal server system and enabling the expected benefits of low power consumption. The lightweight On-Premise LLM is suitable for personal information protection and security and can reduce learning execution time and response latency [9].

This study designed and implemented an architecture for developing a lightweight LLM using knowledge distillation techniques to streamline the creation and processing of public documents used in school administration. The goal was to verify whether the implemented On-Premise LLM could be applied to an actual school administration. The proposed OP-LLM-SA (On-Premise Large Language Model for School Administration) is configured to suit the CPU server environment of schools and enables the creation and processing of official documents in real time within the internal network environment, through data mining based on document summarization, data extraction, and generation through learning existing school administrative documents.

This study contributes with the following:

1. The design and implementation of a Knowledge-Distilled LLM architecture using actual datasets.
2. The verification of the applicability of the proposed model for improving the efficiency of administrative document processing in public institutions (schools, district offices, neighborhood offices, etc.) in an on-premise environment.

This paper is organized as follows. Section 2 reviews text mining technology and knowledge distillation-based compression technology in LLM document processing, including background research. Section 3 presents the design of the architecture of the proposed OP-LLM-SA model. Section 4 presents an evaluation of the implemented model's performance. Section 5 presents the conclusions and future research, focusing on the technical implications of the proposed model and directions for improvement.

Existing research on LLMs in the field of education has mainly focused on educational content; this study is unique in that it specializes in a public document creation system for school administration. In particular, among various administrative tasks, this study focuses on public document processing, which is the most burdensome task for teachers; it is expected that teachers will respond positively to an automation system for this. The proposed OP-LLM-SA model emphasizes practicality by reflecting the legal requirements for school administration data (such as establishing and strengthening personal information protection management systems) while presenting measures to optimize data generation speed and resources. In addition, as a model specialized for public document processing

systems, it is expected to pave the way for AI-based administration solutions across other public institutions, such as district offices and neighborhood offices.

2. Related Work

2.1. Text Mining Technology Trends and Application Cases

Text mining refers to the extraction of required information from large amounts of text data. Text mining technology consists of text preprocessing and feature extraction stages. First, unstructured data is preprocessed through tokenization, stop word removal, stemming, headword extraction, and text normalization, and then, the features are extracted using various technologies. Text mining technology is a field of Natural Language Processing (NLP) that has proven its applicability across various domains with the advancement of machine learning and deep learning technologies [10].

O'Mara-Eves et al. [11] utilized text mining in the research paper domain to automatically filter out necessary references, thereby minimizing the time required to read papers. Gupta et al. [12] analyzed how text mining is utilized in the banking and securities fields, verifying its usefulness in financial data analysis. Gupta et al. [13] confirmed that a BERT-based specialized model can accurately find information in materials engineering papers. Taha et al. [14] analyzed various application cases focusing on core technologies in the field of text mining.

In South Korea, research applying text mining technology to the public sector is also actively underway. Shin et al. [15] applied text mining technology to analyze public data released by local governments in various ways, analyzing the types of local governments, data provision formats, and data characteristics. Through this, they confirmed the need for data openness in various areas that reflect local policies and demands. In addition, Han, S. [16] analyzed news articles related to public records using text mining, and Lee, J.-S. [17] conducted a conceptual expansion study of public design research through text mining.

While most of these studies introduce text mining technology and focus on cases of extracting key information required in various fields (science, finance, public services, etc.), this study is distinguished in that it verifies the practical applicability of text mining technology for the automation and efficiency of school administration documents. In addition, by utilizing an on-premise environment, this study can address the issues of privacy and security that need to be resolved in text mining technology. Table 1 summarizes the main contents of previous studies and this study.

2.2. Knowledge Distillation-Based LLM Lightweight Technology

Large Language Models (LLMs) are efficiently used in Natural Language Processing (NLP) and generative AI, but their large model size, resulting from a large number of parameters, raises issues with the need for large computing space and energy efficiency. A representative technique for solving this problem is Knowledge Distillation (KD). KD extracts meaningful data from a neural network model with a large number of parameters to train a smaller, more efficient model. In other words, it is a method of transferring knowledge from a "teacher" model with a large number of parameters and high performance to a "student" model with a smaller number of parameters. Phuong, M. [18] analyzed the working principles and theoretical basis of the KD technique and mathematically proved that the student model can be trained reliably when soft labels are used.

Mansourian, A.M. [19] emphasized that KD is an essential technology for compressing LLMs and foundation models, and summarized the main techniques of KD technology, such as logit-based, feature-based, attention-based, and self-distillation. They also mention

that knowledge distillation is being applied and developed in various fields such as 3D data, cross-modal learning, and improving adversarial robustness.

Table 1. Comparison of previous studies on text mining technology.

Paper (Source)	Field of Study	Applied Method	Main Objective and Outcome
O'Mara-Eves et al. (2015) [11]	Systematic Literature Review (SR)	Text mining-based literature selection techniques	Expected improvement in the quality of grammatical particles in research papers and reduction in review time
Gupta et al. (2020) [12]	Financial Data Analysis	Various text mining techniques	Evaluation of the applicability of text mining in the financial industry
Gupta et al. (2022) [13]	Materials Engineering Paper Analysis	Domain-specific BERT model (MatSciBERT)	Improvement in data extraction accuracy in materials engineering papers
Taha et al. (2024) [14]	Text Classification Technology	Comprehensive review and experimental analysis of text classification algorithms	Latest text classification techniques and application examples
Shin et al. (2021) [15]	Local Government Public Data	Text mining (keyword frequency analysis, topic analysis)	Understanding the status and characteristics of data openness
Han, S. (2025) [16]	Public Records and News Articles	Keyword analysis, topic modeling	Analysis of major issues and social perceptions in news articles
Lee, J.-S.; Jung, J.-H. (2025) [17]	Public Design Research	Frequency analysis, keyword network analysis	Analysis of the relationship between main concepts and keywords in public design research
This paper	Educational Administration Document Processing	Text mining-based knowledge-distilled LLM	Automation of school administration documents and establishment of an efficient system

Wang et al. [20] explored various cases of KD and emphasized that diversifying data within KD or synthesizing new data is crucial for improving distillation performance. They also argue that, to observe the effectiveness of knowledge distillation, it is necessary to maintain model reliability while aiming for efficiency, transparency, and ethics. Additionally, promising fields such as self-alignment and multi-modal LLMs are mentioned as potential ways to improve KD model performance.

Yang et al. [21] analyzed knowledge distillation algorithms for Large Language Models (LLMs) from three perspectives: methods, evaluation, and application.

To clarify this method, we classify it into white-box KD and black-box KD, with the latter including two distinct types. We also delve into distillation and evaluation methods in the LLM domain and present application examples in healthcare, education, and law. Gu, Y. [22] proposed Mini LLM, which leverages knowledge distillation techniques to maintain LLM functionality in a lightweight model.

While existing KD methods have been widely used to analyze white-box or black-box models, research on effectively transferring small, efficient models into white-box LLMs remains limited. Gu, Y. [22] proposed a method using reverse Kull back-Leibler (KL) divergence to prevent the student model from learning in incorrect areas (overconfidence), enabling the creation of small yet high-quality student LLMs. By transferring white-box LLM functions to a lightweight student model (Mini LLM), they suggest the possibility

of supporting high-quality, high-performance LLM services even in resource-constrained on-premise environments.

Based on these previous studies, this study designed a proposed model that systematizes a knowledge distillation-based LLM suitable for on-premise environments and presents its strategy.

2.3. On-Premise AI System Implementation Technology

AI technology is being used across various industries and is provided in a general cloud environment. However, there is a need for stronger control over learning data and processes, and restrictions may arise due to personal information, so on-premise environments are gaining attention as a core technology. Fortuna, C. [23] presented a plan for building AI services based on an analysis of technology stacks and automation tools for building AI services in on-premise environments. Tachu, E.A. [24] verified the correlation between the flexibility of on-premise infrastructure and the flexibility of cloud infrastructure and presented a model application strategy for modular design.

Recently, various studies on how on-premise infrastructure should be linked with AI technology have been conducted. Pillai, P. [25] analyzed strategic selection factors based on the advantages and disadvantages of on-premise data warehousing and cloud-based warehousing in the financial industry, and Luka, C. [26] proposed a hybrid integration model that combines existing on-premise legacy systems and cloud services.

Gautam [27] presents data storage architectures and design considerations applicable to cloud, hybrid, and on-premise environments for optimizing system performance in the fields of artificial intelligence (AI), generative AI, and retrieval-augmented generation (RAG), which rely heavily on massive amounts of data. Particularly, the importance of data storage in system optimization is emphasized, and the necessity of designing an architecture that considers data storage even in on-premise environments is discussed. Based on these prior studies, this study aims to design a knowledge distillation-based LLM architecture and analyze its practical applicability by constructing an optimal on-premise AI system.

3. Methodology

3.1. Design of the Proposed Model OP-LLM-SA

The proposed model is named OP-LLM-SA (On-Premise Large Language Model for School Administration), as it is used in school administration. The proposed model extracts information from a high-performance teacher model and uses it to effectively train a small student model, as shown in Figure 1.

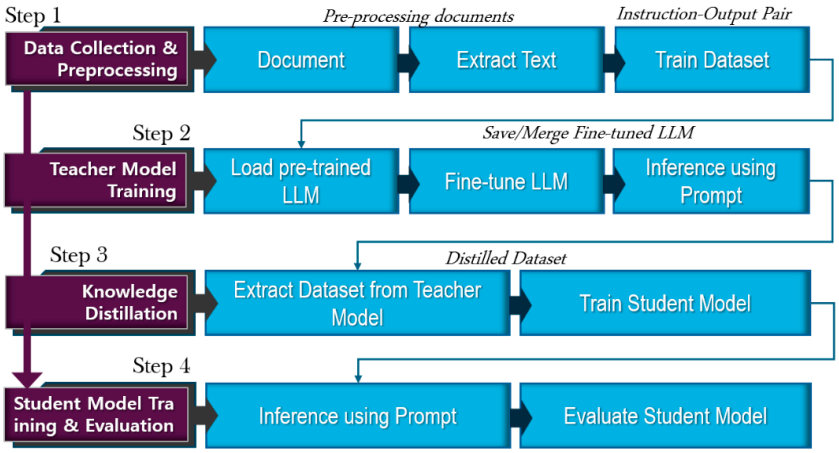

Figure 1. OP-LLM-SA architecture.

It was designed to consist of four stages so that inference would be possible even on computers with CPUs (average memory usage within 4.5 GB), by extracting specific knowledge from a 70B-class Teacher LLM and training Student LLM between 1B and 3B. The four stages are as follows: (1) Data Collection and Preprocessing, in which meaningful text is extracted from various collected documents (Extract Text) and unnecessary grammatical particles and emphatic phrases are preprocessed and removed. The preprocessing stage is then completed while reconstructing the configured dataset. (2) SFT (Supervised Fine-tuning) of the Teacher Model, which is performed to create learning data in the form of Instruction–Output pairs. This is the key to determining the quality of fine-tuned data. (3) Storage of the Instruction–Output pair generated through the teacher model as a Distilled Dataset based on Knowledge Distillation. Multiple response examples are collected through the teacher model and converted for use in training the student model. (4) Student Model Training and Evaluation, in which the lightweight student model is trained, and its performance is verified.

3.2. Knowledge Distillation Pipeline

The knowledge distillation pipeline procedure for the proposed model, OP-LLM-SA, is presented in Figure 2. The pipeline was designed for this study by referencing knowledge distillation research [28,29].

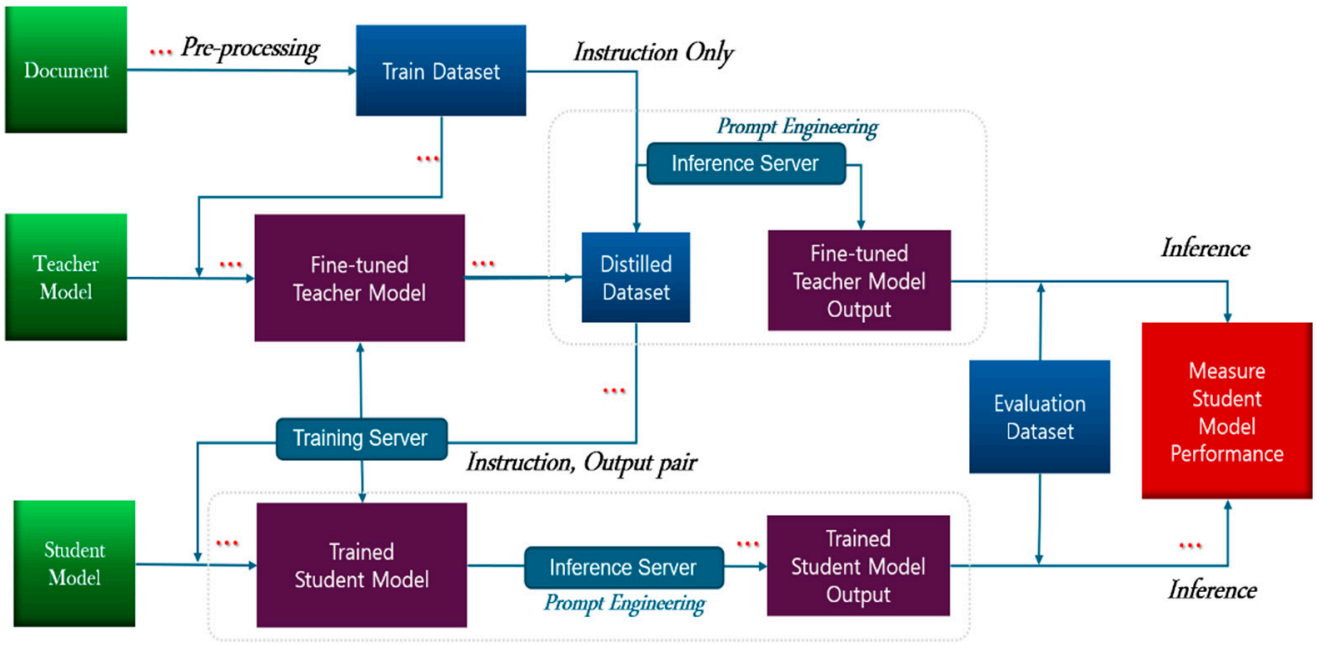

Figure 2. Knowledge distillation pipeline [28,29].

In the proposed model, documents are first organized to create training data. A (large) teacher model with exemplary answers is built, and then a (small) student model learns from this model. The goal of the proposed (small) student model is to learn only the necessary parts lightly, conserving speed and memory while still achieving excellent performance.

Table 2 shows an overview of the knowledge distillation pipeline.

Table 2. Knowledge distillation pipeline overview [28–31].

Block	Description	Key Outputs/Inputs
Administrative Document	Original PDF documents from real-world administrative use cases	Raw text
Preprocessing	OCR-based parsing and segmentation of questions and answers	Cleaned text
Dataset	(a) Instruction–Output pairs; (b) instruction-only prompts	(a) Teacher SFT, KD training and evaluation data
Fine-tuned Teacher	LoRA-based or locally SFT-completed model	High-quality responses
Distilled Dataset	Instruction–Output pairs generated by the teacher model	Student model training data
Student Model	Parameter-optimized lightweight model (e.g., 1B~3B)	Model capacity
Trained Student	Sequential KD-trained model	Final output
Evaluation Dataset	Real-world public query samples	Evaluation metrics (ROUGE, BERTScore)
Measure Performance	Comparative analysis between teacher and student models	Performance evaluation report (quantitative + qualitative)

First, School Administration documents are converted to PDF files using HANCOM 2002. Images are excluded, and during text extraction, unnecessary grammatical particles and emphatic expressions are removed through a preprocessing step. Next, knowledge distillation is performed. Knowledge distillation involves a student model learning from a teacher model to increase data accuracy while minimizing parameters [30].

The teacher model obtained from the constructed dataset undergoes SFT (Supervised Fine-Tuning), during which an appropriate amount of data is used to prevent the teacher model from collapsing. This can potentially happen when fine-tuning to fit the model to the data causes it to learn differently from the original model.

To prevent GPU memory shortage errors during teacher model training and ensure training efficiency, the model is trained using the LoRA (Low-Rank Adaptation) technique. After training, LoRA and the teacher model are merged to create a fine-tuned model that can be used without an internet connection, since the base model is required when using the LoRA model for LLM Inference.

Based on the merged model of the LoRA model (the fine-tuned result of the teacher model) and the original teacher model, we construct an inference prompt to train the student model using the fine-tuned teacher model [31].

This prompt consists solely of instructions without output and is designed to generate various academic administrative documents. The actual knowledge distillation process consists of extracting the dataset from the teacher model. We query the fine-tuned teacher model step-by-step with the inference prompt to obtain actual outputs, which are then used as the training dataset for the student model. The student model uses a low-parameter model that can run sufficiently even on computers or servers equipped with low-specification GPUs. To achieve the maximum effect with minimal data, allowing even low-specification models to work, we limit the number of documents to 80. The student model is trained using distilled prompts, and its outputs are used to construct instruction–output data pairs. Finally, an evaluation dataset is prepared to analyze the

trained student model's performance. This dataset is acquired through inference results from each model, enabling performance analysis.

4. Performance Evaluation of the Implementation Model

4.1. Experimental Environment

By separating the learning server and inference server, the learning–verification–application pipeline was separated to enable operation, enabling parallel research and time optimization, as shown in Table 3. The model learning server optimized model learning through large-scale parallel processing and GPU-to-GPU computation distribution, while the inference server used a single high-performance GPU with relatively low resource intensity, demonstrating the possibility of model compression and real-time response system development.

Table 3. Model inference environment.

	Specifications
Model Inference Environment	✓ Processor (CPU): AMD EPYC™ Genoa 9654 (96 cores/192 threads) (AMD, Sunnyvale, CA, USA) ✓ Memory (RAM): Samsung DDR5 Registered ECC 64 GB × 8 units (Total 512 GB) (Samsung, Suwon-si, Republic of Korea) ✓ Graphics Processing Unit (GPU): NVIDIA RTX 4090 24 GB × 1 unit (NVIDIA, Santa Clara, CA, USA)

4.2. Text Mining Performance Evaluation

In order to quantitatively evaluate the quality of the administrative document outputs generated by the proposed OP-LLM-SA model, we analyzed the data extraction performance based on the generated text [32]. The evaluation of text mining performance is based on 80 official documents. This number was chosen to maximize effectiveness with the minimum number of documents, using a diverse set of 80 official documents to verify feasibility on low-specification computers in an on-premise environment. While this may be considered a relatively small sample, most official documents consist primarily of administrative terminology and are within approximately 500 characters. Therefore, 45 documents were initially selected through preliminary review by six teachers and staff members who regularly draft and utilize official documents. As this study is text-based, documents containing non-textual data such as images or graphs were excluded, resulting in the final selection of 80 documents.

The evaluation items were calculated using four indicators, including token accuracy, completed sentence rate, sentence naturalness, and format suitability, by comparing and analyzing the output documents generated by the model with the original administrative documents and formal appropriateness, as shown in Table 4.

Most of the results showed a high degree of consistency with the original text, demonstrating the model's usability.

- Token Accuracy: Measures word- and phrase-level similarity between the original text and the generated administrative document.

Table 4. Text mining performance evaluation results.

Metric	Percentage (%)	Description
Token Accuracy [32]	92.36%	The percentage of tokens that are identical to or semantically consistent with the original text.
Completed Sentence Rate	97.19%	The percentage of the model's output that matches the sentence-ending structure, relative to the original text (100%).
Sentence Naturalness [33]	99.10%	The fluency score based on the language model's Perplexity, expressed as a percentage converted from the original text (100%).
Format Suitability	92.85%	The reproduction rate of official and administrative document format elements (e.g., title, item, label, date, and amount notation), as a percentage relative to the original text.

It reflects the frequency with which the generated token with the highest probability matches the correct answer, as follows [34].

$$Token\ accuracy = \frac{1}{N}\sum_{i=1}^{N} 1(argmaxP(t|context_i) = t_i^{ref}) \quad (1)$$

where N is the total number of predictions; $argmaxP(t|context_i)$ is the prediction token with the highest probability; t_i^{ref} is the actual correct answer token; and $1(\ldots)$ is an indicator function that returns 1 or 0 to indicate a match.

To evaluate Token Accuracy, titles and content were extracted from 80 public documents, and the student LLM performed inference. Text preprocessing and few-shot prompting techniques were applied to ensure output consistency. Accuracy was measured at 92.36%, with most core vocabulary and syntax matching the original text.

- Sentence Naturalness: This is an indicator that determines how natural the generated administrative documents are, and it was measured at 99.10%, which is almost identical to the original text in terms of grammar and fluency [33].
- Completed Sentence Rate: This metric measures the average sentence length and the percentage of complete sentences in generated administrative documents. Sentence completeness was measured using a GPT model, yielding an average sentence length of 15.68 characters and a complete sentence rate of 97.19% when compared to existing official documents.
- Format Conformity: This metric measures the degree of adherence to official document formatting standards. Measured using the GPT model, it achieved 92.85%, indicating a high rate of format element reproduction.

4.3. System Efficiency Evaluation

The efficiency of the on-premise system was evaluated by measuring the average CPU, RAM, and GPU usage during teacher model learning and inference, and student model learning and inference, as shown in Table 5.

Table 5. On-premise system efficiency evaluation.

Average Usage (%)	Teacher Model Training	Teacher Model KD	Student Model Training	Student Model Inference
CPU	0.34%	2.86%	0.06%	0.07%
RAM	3.15%	3.16%	3.2%	3.38%
GPU	64.51% (126,804 MB)	36.92% (72,575.84 MB)	10.80% (4709.3 MB)	29.28% (4583.11 MB)

The reason for measuring CPU, RAM, and GPU usage was to determine whether bottlenecks occurred on computers used for actual school administration and office work, which are not high-performance servers but rather general office computers, when generating official documents using the student model.

The evaluation results showed that meaningful results could be obtained with the current PC environment in elementary, middle, and high schools in South Korea. In particular, the GPU memory usage in the student model's inference task environment was measured at approximately 4.5 GB, confirming that the model can run sufficiently even on general office computers equipped with GPUs with small memory capacities.

4.4. LLM Performance Evaluation

To evaluate the performance of the proposed OP-LLM-SA model, we conducted a performance analysis using comparative indicators to determine whether it was properly trained from a knowledge distillation perspective and whether its performance could be effectively manifested. We compared the fine-tuned teacher model, the lightweight vanilla student model, and the proposed OP-LLM-SA student model. First, the fine-tuned teacher model is an ideal teacher model generated through data learning, while the vanilla student model is a general-state student model. The OP-LLM-SA model is a small-scale model but has been trained using the proposed architecture with an effective framework. The appropriateness of responses for each question is compared as follows.

BLEU (Bilingual Evaluation Understudy) evaluates how similar a student model is to a teacher model [35]. BLEU analyzes precision based on sentence length as follows.

$$BLEU = BP \times \exp\left(\sum_{n=1}^{N} w_n log p_n\right) \tag{2}$$

where p_n is the n-gram precision; w_n is the weight of each n-gram precision; and BP is the Brevity Penalty—a penalty imposed if the translation is too short.

$$BP = \begin{cases} 1 & , \ c > r \\ \exp(1 - \frac{r}{c}), & c \leq r \end{cases} \tag{3}$$

where c is the length of the candidate translation and r is the length of the reference translation.

ROUGE-1 (Recall-Oriented Understudy for Gisting Evaluation) is a simple measure of text similarity based on the unigram (word-level) recall rate, while ROUGE-L extends this by reflecting whether the word order within sentences is preserved [36].

$$ROUGE - 1 = \frac{\sum_{w \in Ref} \min(Count_{Cand}(w), Count_{Ref}(w))}{\sum_{w \in Ref} Count_{Ref}(w)} \tag{4}$$

where $Count_{Cand}(w)$ is the number of times a word appears in the candidate summary or translation and $Count_{Ref}(w)$ is the number of times a word appears in the reference summary or translation.

ROUGE-L evaluates not only the word overlap but also the degree to which the word order within sentences is preserved. Therefore, ROUGE-L complements ROUGE-1 by enabling the evaluation of both word coverage (ROUGE-1) and sentence-level similarity (ROUGE-L).

$$ROUGE - L_{rec} = \frac{LCS(X,Y)}{|Y|} \ ROUGE - L_{prec} = \frac{LCS(X,Y)}{|X|} \tag{5}$$

$$ROUGE - L_f = \frac{(1+\beta^2) \times ROUGE - L_{rec} \times ROUGE - L_{prec}}{ROUGE - L_{rec} + \beta^2 \times ROUGE - L_{prec}} \quad (6)$$

where X is the candidate summary; Y is the reference summary; $LCS(X, Y)$ is the length of the Longest Common Subsequence (LCS) between two sentences; and β is a weighting scheme that places greater importance on recall than on precision (commonly set to around 1.2).

BERT_Score evaluates sentence-level semantic similarity, measuring completeness through understanding the meaning between sentences [37].

(1) Each token is embedded using a pre-trained language model (e.g., BERT, RoBERTa) to vectorize it. (2) The cosine similarity matrix is calculated. (3) Precision and Recall are calculated as follows.

$$F1 = \frac{2PR}{P+R} \quad (7)$$

where P is Precision, computed as the average similarity between each token in the candidate sentence and its most similar token in the reference sentence, and R is Recall, computed as the average similarity between each token in the reference sentence and its most similar token in the candidate sentence.

The results of evaluating the performance of the proposed model based on the original document are shown in Table 6. The proposed OP-LLM-SA model achieved BLEU 97.20, ROUGE-L 99.04, and BERT_Score 98.29, confirming its overall high performance compared to vanilla student. In particular, BLEU and ROUGE-L showed figures close to those of the teacher model, indicating excellent token accuracy and structural reproducibility, suggesting suitability for official document creation. All are statistically significantly superior within the 3B group ($\Delta \geq +2$ percentage points or more).

Table 6. LLM performance comparison results (unit: %).

Model	#Params	Method	BLEU	ROUGE-1	ROUGE-L	BERT_Score
llama-3.2-Korean-Blossom	70B	Fine-tuned teacher	97.20	99.05	99.04	98.29
	3B	Vanilla Student	87.50	95.04	94.68	96.09
	3B	OP-LLM-SA	94.30	98.09	98.18	98.55
llama-3.2-instruct	3B	Fine-tuned teacher	58.54	69.42	68.41	86.78
	1B	Vanilla Student	43.57	76.15	73.73	86.36
	1B	OP-LLM-SA	45.31	77.47	76.04	86.94

The Llama-3.2-instruct 3B teacher model showed lower performance with BLEU 58.54 and ROUGE-L 68.41. This highlights that domain-specific training data is a critical factor in model performance for the Korean language domain.

To assess the qualitative effectiveness of the generated official documents, a qualitative evaluation (FGI) was conducted with six current teachers who draft and handle administrative documents. Identifying factors influencing teachers' acceptance attitudes and intentions toward IT application in educational settings is crucial for enhancing educational quality and systematizing administration [38].

The following image shows an official document generated based on the original text. Six current teachers responsible for drafting and managing administrative documents conducted a qualitative evaluation of this document, and the result is shown in Figure 3.

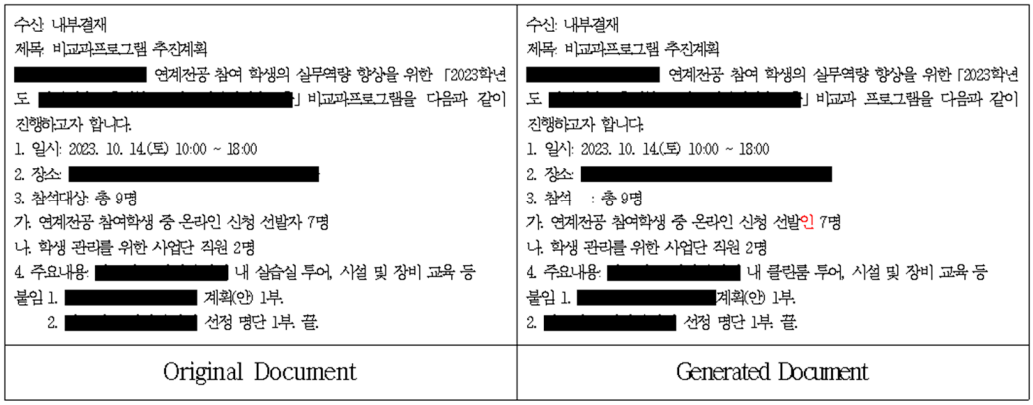

Figure 3. Comparison of the original and model-generated documents.

They assessed structural appropriateness, content completeness, and Sentence Naturalness according to the quantitative evaluation format above. All evaluators showed that the content generated by the proposed model was directly usable. They assessed it as being nearly identical in format to existing administrative documents, to the extent that drafting and editing official documents would be unnecessary.

They also noted the advantage of operating in an on-premise environment, allowing free use of School Administration-specific terminology. However, they expressed regret that only text is currently supported. Accordingly, they suggested improvements to enable document generation beyond simple text, including images and graphs.

- The format and structure of the proposed model are suitable, and the ratio of completed sentences is also excellent.
- While there are concerns about the leakage of terms used in schools when using ChatGPT4o, the proposed model appears ready for immediate use as it eliminates the risk of external leakage.
- It is regrettable that only text generation is possible. Please enable the use of photos or graphs.

5. Conclusions

This study designed and implemented a knowledge-distilled large language model (Knowledge-Distilled LLM) architecture for the efficient management of school administrative documents. The proposed OP-LLM-SA model extracted data from a high-performance teacher model and effectively trained a compact model using this knowledge.

We preprocessed 80 official document files used in actual School Administration and performed knowledge distillation through fine-tuning of the teacher model. The system was completed with a student model trained on the results generated through this process. Text mining performance evaluation showed very high consistency with the original text. With a token accuracy of 92.36% and a complete sentence rate of 97.19%, the generated administrative documents were confirmed to be immediately usable in terms of grammatical accuracy and fluency.

Furthermore, the server for model training optimized this process through distributed computing. The inference server was confirmed to use relatively few resources, approximately 4.5 GB of GPU memory, proving its applicability even on general computers used in schools.

Additionally, meaningful results were obtained from the performance comparison. We compared a fine-tuned teacher model, a lightweight vanilla student model, and the proposed OP-LLM-SA student model. The fine-tuned teacher model is an ideal teacher model generated through data learning, while the vanilla student model is a standard

student model. Despite being a small model, the OP-LLM-SA model demonstrated a significant improvement in data learning efficiency when evaluated using the proposed effective system architecture. The proposed model achieved the following scores, demonstrating overall higher performance and practical applicability compared to the standard student model: BLEU 97.20, ROUGE-L 99.04, and BERT_Score 98.29. This suggests the proposed model is highly suitable for the reproduction and generation of official documents through directly learning the teacher model's output sequence.

This study is significant as it verifies the applicability of the proposed model in an on-premise environment to enhance the efficiency of administrative document processing in public institutions (schools, district offices, neighborhood offices, etc.). Its significance is further amplified as it represents the first attempt in a School Administration setting beyond the existing medical and public sectors.

However, this study has the following limitations.

The proposed model is specialized for the Korean domain, limiting its generalization due to potential biases and differences in administrative documents outside Korea. Specifically, it was confirmed that using a model with training data optimized for the specific domain is a decisive factor in performance during document generation. For example, Llama-3.2-instruct 3B Teacher showed a relatively low performance, with the following scores: BLEU 58.54 and ROUGE-L 68.41. While the performance of the student model in the same family was improved, the performance gap compared to Blossom family models was found to be quite significant.

Furthermore, when verifying the model's feasibility on low-spec computers in on-premise environments, only 80 official documents were used. However, judging quantitative metrics based on just 80 official documents has limitations. This study can be viewed as foundational research that demonstrates that accurate extraction of official documents is possible even with limited data utilization on low-spec computers.

Furthermore, this research was conducted under the premise that personal information is unconditionally and perfectly protected within on-premise environments in South Korea, where personal information protection systems in public institutions are extremely stringent. However, the potential risks of unexpected data leaks and mitigation strategies for these risks were not sufficiently explored.

The proposed model focused on performance evaluation and analysis, emphasizing usability on low-performance computers and text mining performance metrics, but did not address real-time processing capabilities through metrics such as speed. However, generative AI boasts extremely fast speeds, meaning that the speed of document generation through training is critically important for commercialization of the proposed model. Therefore, in future research, there are plans to perform a more comprehensive study by setting document generation conditions using various foreign languages, verifying practicality by considering the processing time during performance evaluations, and applying data security and guidelines for handling public documents when personal information is generated.

Funding: This research received no external funding.

Data Availability Statement: Data will be made available on request.

Acknowledgments: During the preparation of this study, the author used ChatGPT-4.0 for the purposes of performance evaluation in the text mining analysis. The author has reviewed and edited the output and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Choi, H.-Y.; Choi, Y. Diagnosis of Public Language Used in Administrative Agencies for Develop 'How to Write Official Documents' Teaching Materials. *J. Lang. Lit.* **2014**, *58*, 53–75. [CrossRef]
2. Cho, Y. Analysis of Teachers' Job Burden and Influencing Factors by Job Area. *J. Educ. Res.* **2024**, *22*, 89–114. [CrossRef]
3. Lee, H.-S. The Administrative Work Burden of Childcare Teachers and Improvement Measures. Master's Thesis. Kookmin University Graduate School of Education, Seoul, Republic of Korea, 2018. Available online: https://www.riss.kr/link?id=T14897125 (accessed on 7 July 2025).
4. Mandvikar, S. Augmenting Intelligent Document Processing (IDP) Workflows with Contemporary Large Language Models (LLMs). *Int. J. Comput. Trends Technol.* **2023**, *71*, 80–91. [CrossRef]
5. Oh, Y. Issues and Prospects of AI Digital English Education: Text Analysis Based on Natural Language Processing. *Korea J. Engl. Lang. Linguist.* **2025**, *25*, 330–366. [CrossRef]
6. Jo, W. The Possibility of Collective Emotion Narrative Research Using Self-Attention. *Korean J. Sociol.* **2024**, *58*, 315–350. [CrossRef]
7. Miranda, M.; Beugnot, G.; Heitz, M.; Vert, J.-P. Preserving Privacy in Large Language Models: A Survey on Current Threats and Solutions. *arXiv* **2024**, arXiv:2408.05212. [CrossRef]
8. Personal Information Protection Act. Article 32-2; Amended 8 March 2023. Available online: https://www.law.go.kr/LSW/lsInfoP.do?lsiSeq=250244&efYd=20230308 (accessed on 19 July 2025).
9. Liu, C.; Zhang, H.; Zhao, K.; Ju, X.; Yang, L. LLMEmbed: Rethinking Lightweight LLM's Genuine Function in Text Classification. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), Bangkok, Thailand, 11–16 August 2024; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 7997–8011. Available online: https://aclanthology.org/2024.acl-long.433.pdf (accessed on 20 April 2025).
10. Aleqabie, H.J.; Sfoq, M.S.; Albeer, R.A.; Abd, E.H. A Review of Text Mining Techniques: Trends, and Applications in Various Domains. *Iraqi J. Comput. Sci. Math.* **2024**, *5*, 9. [CrossRef]
11. O'Mara-Eves, A.; Thomas, J.; McNaught, J.; Miwa, M.; Ananiadou, S. Using Text Mining for Study Identification in Systematic Reviews: A Systematic Review of Current Approaches. *Syst. Rev.* **2015**, *4*, 5. [CrossRef] [PubMed]
12. Gupta, A.; Lamba, H.; Kumaraguru, P.; Joshi, A. Comprehensive Review of Text Mining Applications in Finance. *Financ. Innov.* **2020**, *6*, 39. [CrossRef]
13. Gupta, T.; Himmelstein, D.S.; Baranzini, S.E.; Greene, C.S. MatSciBERT: A Materials Domain Language Model for Text Mining and Information Extraction. *npj Comput. Mater.* **2022**, *8*, 217. [CrossRef]
14. Taha, K.; Yoo, P.D.; Yeun, C.; Homouz, D.; Taha, A. A Comprehensive Survey of Text Classification Techniques and Their Research Applications: Observational and Experimental Insights. *Comput. Sci. Rev.* **2024**, *54*, 100664. [CrossRef]
15. Shin, S.-Y.; Choi, H.; Kim, D. Analyzing Contents of Open Data in Korean Local Governments: A Text-Mining Approach. *Korean J. Public Adm.* **2021**, *30*, 129–171.
16. Han, S.; Lee, H. An Analysis of News Articles Related to Public Records Using Text Mining. *J. Korean Soc. Archival Rec. Manag.* **2025**, *25*, 103–127. [CrossRef]
17. Lee, J.-S.; Jung, J.-H. A Study on the Conceptual Expansion of Public Design Research through Text Mining. *J. Public Des.* **2025**, *16*, 57–66. [CrossRef]
18. Phuong, M.; Lampert, C.H. Towards Understanding Knowledge Distillation. In Proceedings of the 36th International Conference on Machine Learning (ICML 2019), Long Beach, CA, USA, 9–15 June 2019; PMLR: Cambridge, MA, USA, 2019; Volume 97, pp. 5142–5151. Available online: https://proceedings.mlr.press/v97/phuong19a.html (accessed on 28 March 2025).
19. Mansourian, A.M.; Saadatfar, H.; Torkamani, M.; Kazemi, H.; Zolbanin, H.M.; Moghaddam, H.A. A Comprehensive Survey on Knowledge Distillation. *arXiv* **2025**, arXiv:2503.12067. [CrossRef]
20. Xu, X.; Li, M.; Tao, C.; Shen, T.; Cheng, R.; Li, J.; Xu, C.; Tao, D.; Zhou, T. A Survey on Knowledge Distillation of Large Language Models. *arXiv* **2024**, arXiv:2402.13116.
21. Yang, C.; Lu, W.; Zhu, Y.; Wang, Y.; Chen, Q.; Gao, C.; Yan, B.; Chen, Y. Survey on Knowledge Distillation for Large Language Models: Methods, Evaluation, and Application. *arXiv* **2024**, arXiv:2407.01885. [CrossRef]
22. Gu, Y.; Dong, L.; Wei, F.; Huang, M. MiniLLM: Knowledge Distillation of Large Language Models. In Proceedings of the International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 5–9 May 2024. Available online: https://openreview.net/forum?id=5h0qf7IBZZ (accessed on 7 June 2025).
23. Fortuna, C.; Mušič, D.; Čerar, G.; Čampa, A.; Kapsalis, P.; Mohorčič, M. On-Premise Artificial Intelligence as a Service for Small and Medium Size Setups. *arXiv* **2022**, arXiv:2210.06956. [CrossRef]
24. Tachu, E. A Quantitative Study of the Relationship between Cloud Flexibility and On-Premise Flexibility. *Issues Inf. Syst.* **2022**, *23*, 214–238.
25. Pillai, P. Cloud vs. On-Premise Data Warehousing: A Strategic Analysis for Financial Institutions. *J. Comput. Sci. Technol. Stud.* **2025**, *7*, 503–513. [CrossRef]

26. Luka, C. Hybrid Integration Model: Integrating On-Premise Legacy Systems with Azure and AI Technologies. *ResearchGate Preprint* **2025**. Available online: https://www.researchgate.net (accessed on 2 June 2025).
27. Gautam, A. Optimizing Data Storage for AI, Generative AI, and Machine Learning: Challenges, Architectures, and Future Direction. *Int. J. Comput. Appl.* **2025**, *186*, 29–33. [CrossRef]
28. Yu, Y.; Kim, N. Text Classification Using Heterogeneous Knowledge Distillation. *J. Korea Soc. Comput. Inf.* **2022**, *27*, 29–41. [CrossRef]
29. Zou, Y.; Xu, Z.; Zhang, Q.; Lin, Z.; Wang, T.; Liu, Z.; Li, D. Few-Shot Learning With Manifold-Enhanced LLM for Handling Anomalous Perception Inputs in Autonomous Driving. *IEEE Trans. Intell. Transp. Syst. Early Access* **2025**, 1–18. [CrossRef]
30. Saxena, P.; Janzen, S.; Maaß, W. Streamlining LLMs: Adaptive Knowledge Distillation for Tailored Language Models. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), Albuquerque, NM, USA, 29 April–4 May 2025; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 448–455. [CrossRef]
31. Fu, Z.; Yang, H.; So, A.M.-C.; Lam, W.; Bing, L.; Collier, N. On the Effectiveness of Parameter-Efficient Fine-Tuning. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington DC, USA, 7–14 February 2023; Volume 37, pp. 12799–12807. [CrossRef]
32. Xia, C.; Xing, C.; Du, J.; Yang, X.; Feng, Y.; Xu, R.; Yin, W.; Xiong, C. FOFO: A Benchmark to Evaluate LLMs' Format-Following Capability. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 11–16 August 2024; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 680–699. [CrossRef]
33. Bahl, L.R.; Jelinek, F.; Mercer, R.L. A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **1983**, *5*, 179–190. [CrossRef] [PubMed]
34. Shlegeris, B.; Roger, F.; Chan, L.; McLean, E. Language Models Are Better Than Humans at Next-Token Prediction. *Trans. Mach. Learn. Res.* **2024**. Available online: https://openreview.net/forum?id=RNsnSLdmV7 (accessed on 13 July 2025).
35. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.-J. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), Stroudsburg, PA, USA, 7–12 July 2002; Association for Computational Linguistics: Philadelphia, PA, USA, 2002; pp. 311–318. [CrossRef]
36. Lin, C.-Y. ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, 25–26 July 2004; Association for Computational Linguistics: Stroudsburg, PA, USA, 2004; pp. 74–81. Available online: https://aclanthology.org/W04-1013.pdf (accessed on 5 July 2025).
37. Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K.Q.; Artzi, Y. BERTScore: Evaluating Text Generation with BERT. In Proceedings of the International Conference on Learning Representations (ICLR 2020), Addis Ababa, Ethiopia, 26–30 April 2020; OpenReview: Alameda, CA, USA, 2020. Available online: https://openreview.net/forum?id=SkeHuCVFDr (accessed on 22 July 2025).
38. Tsai, C.-C.; Lin, Y.-X.; Lee, C.-I. Enhancing education quality: Exploring teachers' attitudes and intentions towards intelligent MR devices. *Eur. J. Educ.* **2024**, *59*, e12692. [CrossRef]